浙 江 农 林 大 学 学 报，2017，**34**(4)：743–750
Journal of Zhejiang A & F University
doi:10.11833/j.issn.2095-0756.2017.04.022

基于网络爬虫的森林经营知识采集系统研建

刘建成，吴保国，陈　栋

（北京林业大学 信息学院，北京 100083）

摘要：针对如何在互联网上准确获取森林经营知识的问题，提出研建森林经营知识采集系统来解决这一问题。在分析森林经营知识采集问题的基础上，设计系统流程、系统模块、数据库，改进网络爬虫规则并加以限定，论述爬虫工作流程和算法。该系统总结分析了森林经营主题网页的特点，通过建立森林经营特征向量对采集内容进行识别，并对森林经营知识去噪处理，智能匹配规则提取知识，使用欧氏距离识别指纹去除重复的森林经营知识。实验结果表明，该系统采集的森林经营知识具有高主题相关度、高准确率、低重复度的特点，满足服务于森林经营决策支持系统的要求。图 7 表 1 参 13

关键词：森林经理学；森林经营知识；知识库；知识采集；网络爬虫

中图分类号：S750　　**文献标志码**：A　　**文章编号**：2095-0756(2017)04-0743-08

Research and construction of web crawler based forest management knowledge collection system

LIU Jiancheng, WU Baoguo, CHEN Dong

(School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China)

Abstract: Accurate Internet access to forest management information can be obtained through the construction of a data collection system for forest management. Based on an analysis of the data collection, system process, system module and database were designed, rules governing web crawlers were improved and delimited, and workflow and algorithm of web crawlers were explored. This system summarized and analyzed the characteristics observed from webpages featuring forest management, and served to identify those collected data contents with an eigenvector of forest management. Information about forest management was also denoised by this system; information was extracted through intelligence match, and repeated information about forest management was eliminated through fingerprint recognition by Euclidean distance. The experiment results indicated that this data collection system for forest management featured high subject relevance, high accuracy, and low repetition rate. Therefore, it can satisfy the need of the forest management decision support system. [Ch, 7 fig. 1 tab. 13 ref.]

Key words: forest management; forest management knowledge; knowledge base; knowledge collection; web crawler

决策支持系统[1]处理问题能力由知识库的知识丰富度决定，如何提升知识丰富度是一个难题。通过网络爬虫采集信息，识别其中的森林经营知识，并进行评价、提取、去重，可以解决这一问题。传统的搜索引擎有强大的网络爬虫，覆盖面广，但分类专业性较差，信息搜索结果不尽如人意[2]，不能准确理解林业词汇。以林业常用名词“小班”为例，百度检索出来的结果绝大多数是幼儿园小班有关的结果，不能满足林业用户的信息检索需求。林业关于信息采集的研究大部分集中在林业主题搜索引擎的研究上，重点研究林业主题搜索引擎的设计、主题爬虫算法、信息源发现方法等算法优化问题[3–7]，但对森

收稿日期：2016-07-14；修回日期：2016-11-04

基金项目：“十二五”国家高技术研究发展计划(“863”计划)项目(2012AA102003)

作者简介：刘建成，博士研究生，从事林业决策支持系统与信息技术研究。E-mail：liujiancheng1018@163.com。通信作者：吴保国，教授、博士生导师，从事林业信息技术研究。E-mail：wubg@bjfu.edu.cn

林经营知识识别、提取等涉及较少。作者通过对主要的森林经营网站进行分析，设计了森林经营知识采集系统的基本工作流程、系统功能模块和数据库，改进了网络爬虫规则，研究森林经营主题爬虫算法、森林经营网页去噪、森林经营知识智能匹配、森林经营知识去重等。

1 森林经营知识采集系统设计

1.1 系统的设计目标与功能

森林经营知识采集系统服务于森林经营决策支持系统，系统目标是通过网络爬虫抓取互联网上与森林经营知识有关的数据，并进行去噪、识别、提取、评价、去重，找出其中可信的森林经营知识，丰富森林经营知识库，提升森林经营决策支持系统的问题处理能力。该系统的建立将提升森林经营决策支持系统的知识丰富度，使森林经营决策支持系统的知识更新与互联网的信息更新同步进行，解决通过人工进行知识获取与整理的效率低、速度慢、信息旧等问题。

森林经营知识采集系统包括初始化模块、存储模块、地址队列生成模块、网页抓取模块、解析模块、森林经营知识过滤模块、森林经营知识提取模块、森林经营知识格式化处理模块等。模块之间协同工作，组成了一个有机的整体，实现了爬虫地址队列生成、网页信息抓取、森林经营知识去噪、森林经营知识识别、森林经营知识提取、森林经营知识评价、森林经营知识去重、森林经营知识存储等功能。

1.2 基本工作流程设计

森林经营知识采集系统的基本工作流程如图 1：①网络爬虫从森林经营主题地址库中获取地址，形成工作队列；②根据爬虫获取到的地址队列抓取地址对应的 HTML(hyper text markup language，超文本标记语言)源码文档，并进行森林经营知识去噪；③根据源码特征智能匹配规则，抽取文本信息特征，建立目标网页向量，抽取链接地址并将相同的地址合并，添加到爬虫工作队列；④将目标网页向量与森林经营特征向量进行计算，识别与森林经营知识无关的网页并过滤；⑤森林经营知识抽取，对抽取的知识进行指纹相似度计算，去除重复知识，对知识进行森林经营主题相关性评价，输出符合要求的森林经营知识存入知识库；⑥对数据库的数据使用森林经营词库进行分词处理，创建索引库，方便检索。

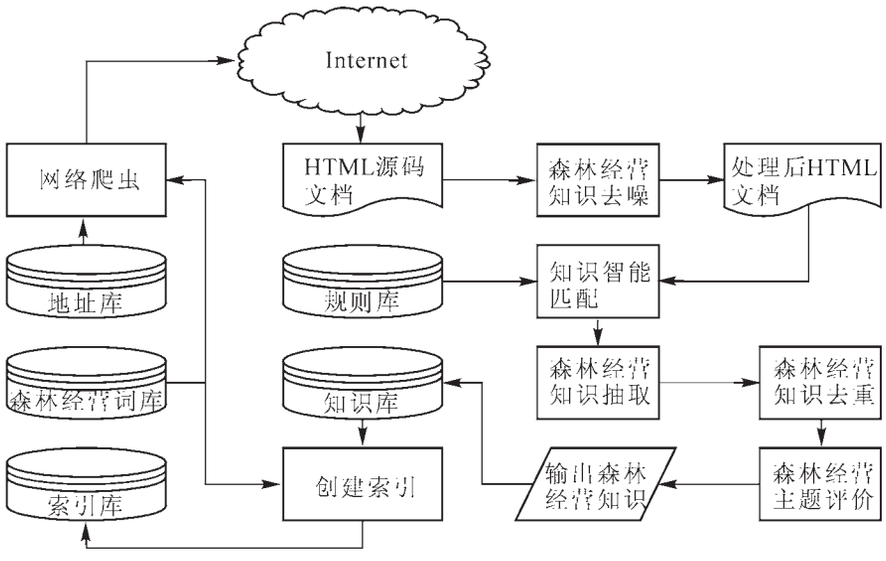

图 1 系统基本工作流程图

Figure 1 Basic working flow chart of the system

1.3 系统功能模块

森林经营知识采集系统包括以下模块：①初始化模块。读取原始地址库中的地址数据，将初始地址数据进行过滤和合并，将初始化数据提交给存储模块进行存储。②存储模块。存储初始化数据、地址队列数据、网页解析数据、提取后的地址数据、提取后的森林经营知识数据。③地址队列生成模块。计算地址数据的权重得分，根据权重得分对地址数据进行处理，生成地址队列。④网页抓取模块。根据爬虫工作的地址队列，利用多线程技术抓取地址在互联网上的对应网页，获取网页文档。⑤解析模块。根据

HTTP协议，将抓取到的网页转译为 HTML 源码，识别源码中的噪音信息并去除，提交给存储模块。⑥森林经营知识过滤模块。根据解析后的网页源码建立向量，与系统中的森林经营知识特征向量进行 Pearson 相关系数计算。判别计算结果与阈值的关系，过滤掉与森林经营主题无关的网页。⑦森林经营知识提取模块。识别网页的源码特征，根据源码特征智能匹配提取规则，提取其中的知识数据和地址数据。⑧格式化处理模块。按照森林经营决策支持系统知识库的知识格式，根据属性分类提取森林经营知识数据并进行格式化处理，创建知识库索引。森林经营知识采集系统功能模块图如图 2 所示。

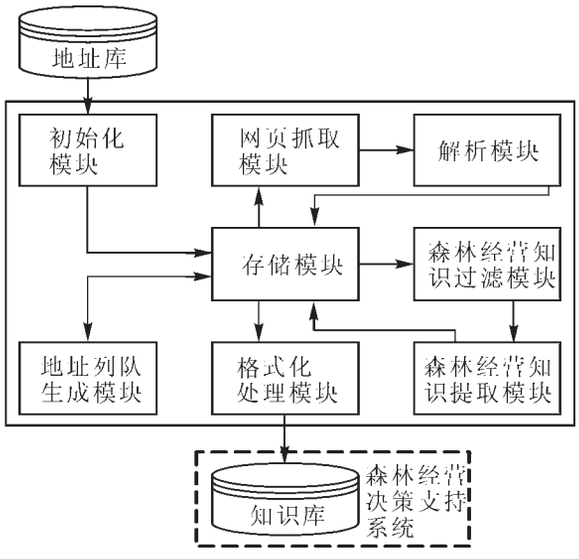

图 2　系统功能模块图

Figure 2　system functional modules diagram

1.4　数据库设计

考虑到数据实体间的关系，本研究设计了管理员表、爬虫表、网站表、规则表、结果表、知识表等 6 个数据表。实体间的联系如图 3 所示。

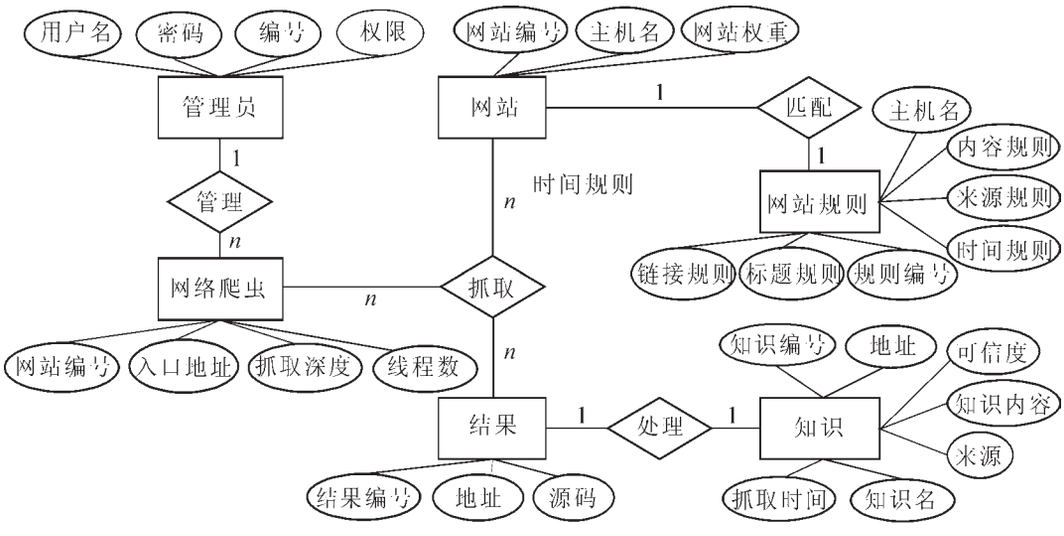

图 3　实体-联系图

Figure 3　Entity-relationship diagram

2　森林经营主题爬虫算法

2.1　网络爬虫规则改进

普通网络爬虫是一个自动提取网页的程序[8]，它提供了一种信息获取的方法。普通网络爬虫一般以一个或多个初始的 URL(uniform resource locator，统一资源定位符)作为入口，重复提取目标网页源码中的 URL 形成队列，并将新的 URL 补充到待爬取的 URL 队列，直至满足系统停止条件为止。普通爬虫并不对内容进行识别，不对链接进行过滤，也不对内容进行分类，仅仅只是宽泛地爬取链接，并获得链接对应的网页内容。森林经营主题爬虫则根据森林经营网页识别算法，过滤与森林经营主题无关的链接，保留与森林经营主题相关的链接，再形成爬虫的待爬取 URL 队列。在抓取的过程中，爬虫对抓取内容进行存储、提取并识别内容是否符合森林经营主题，根据内容辨识从属的分类，并建立索引方便检索。此外，森林经营主题爬虫还根据行政等级划分对网站进行 Rank 评级，保证了内容可信度；添加森林经营专有词库，准确识别森林经营专有名词；建立森林经营特征向量过滤抓取内容。

2.2 森林经营主题爬虫

2.2.1 *爬虫限定规则* 本研究按照网络爬虫[9]的一般系统结构和实现技术，结合森林经营管理知识的特点和网络爬虫模块的设计需要，采用常用主题爬虫技术与深度优先爬虫进行结合，并增加了爬虫抓取的限定规则。森林经营知识具有一定的专业性，对信息来源可信度要求较高。中国林业承担的公益性任务居多，林业发展和林业研究主要依靠政府的财政投入和政策引导。就林业行业的整体权威性来说，政府、科研单位、高校高于企业和社会机构，中央单位高于地方单位。因此，本研究以权威性作为依据，爬虫并不宽泛对互联网的数据进行爬取，按照国家、省(自治区或直辖市)、市、县(区)的行政等级划分对网站进行排队评级，行政等级越高的单位，网站权重越高，可信度越大。在网络爬虫的设计上，充分考虑到森林经营知识的可靠性要求，只对地址库中存储的主机名内的链接进行处理。在工作过程中，系统对采集链接的锚文本进行主题相关度计算，只有符合主题需求的链接才会添加到工作队列。此外，爬虫根据网站的排队评级确定地址得分，通过得分确定工作队列的采集顺序。本研究采用 PanGuAnalyzer 作为系统切词器，对网站的页面信息进行处理。在非林业行业网页中，都可以发现一些林业相关词汇的广告信息，如幼儿园小班面授、幼儿园小班授课等，但是林业网站中常常会有如：林小班、造林小班、森林经营管理小班等词汇。如果不对这类词汇进行区分，赋予其森林经营领域的特定含义进行识别，爬虫可能会识别为广告信息进行过滤。本研究在系统词库的基础上，添加了林业专有词汇，如造林模式、宜林地、迹地拨交、小班、林班、细班、林相图、褐斑病等近 2 000 条以满足需要。涵盖了森林培育与管护、森林经营与决策、森林病虫害、造林树种、林木收获利用等方面的常用词汇。为了避免树种拉丁名信息被当成普通英文字母或者语法错误、拼写错误的英文单词，本研究还在词库中添加了中国主要造林树种的拉丁学名。

2.2.2 *网络爬虫算法* 本研究的网络爬虫工作流程如图 4 所示。其算法如下：第 1 步，预读数据，对 URL 队列进行优先级得分计算。第 2 步，根据 URL 优先级得分，重新排序。第 3 步根据 URL 获取对应网页源码文档。第 4 步提取网页源码文档 URL 链接组进入 Todo 工作队列。第 5 步，获取 Todo 工作队列的第 i 条数据的 URL 即 Todo[i–1].Url，判别 URL 是否是在 URL 索引库的主机名范围内。第 6 步如果是，提取页面信息并对信息进行格式化处理等进一步操作，i+=1，执行第 3 步；如果否，跳过页面信息处理，i+=1，执行第 5 步。第 7 步，判断 Todo 队列是否还有 URL 未处理，如果有，i+=1，执行第 5 步；如果没有，执行跳转到第 8 步。第 8 步工作结束。通过只对站内链接处理的方法，保证了信息的采集来源都是索引数据库中提供的网站，提高了信息来源的可靠性。

2.2.3 *抓取内容过滤* 一个森林经营主题网页通常由树种信息、林学知识、作业模式信息、经营模式信息、技术措施信息和森林病虫害信息这几类构成，而这些信息一般混合了种名、拉丁名、树种类型、树种特性、经营作业技术、病害信息、虫害信息等特性。爬虫采集到的数据不一定是森林经营知识，因此要对爬虫采集的知识进行过滤，只保留森林经营相关知识。本研究根据 VSM(vector space model，向量空间模型)建立森林经营 n 维特征向量(种名，拉丁名，树种类型，特性，采种技术，…)来辨别目标网页是否与森林经营知识有主题关联，特征向量记为 V^T=[(t_1，ω_1)，(t_2，ω_2)，…，(t_n，ω_n)]，其中 ω_i(i=1，2，3，…，n)表示不同属性对应的权值，由 TF-IDF[10]方法确定。

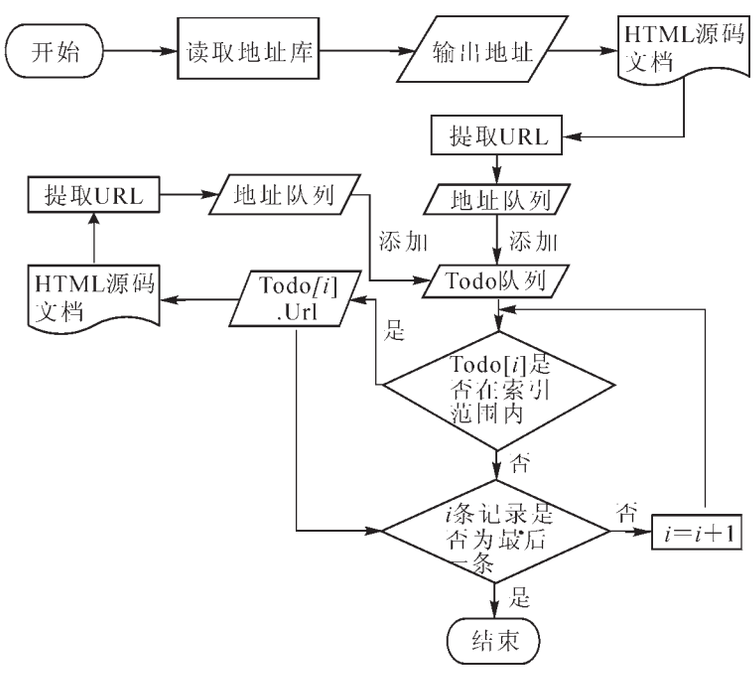

图 4 爬虫算法流程图

Figure 4 Flow chart of the crawler algorithm

3　系统实现关键技术

3.1　森林经营网页去噪

对主要的林业行业网站进行分析可以发现，一个森林经营主题网页的 HTML 源码通常包含 head 和 body 等 2 个部分。网络爬虫对地址队列进行采集处理，通过.NET 的 Stream 类和 WebClient 类的实例化操作来获取地址队列对应的森林经营网页文档的 HTML 源码。对爬虫采集到的森林经营主题网页绘制树形节点结构，其树形节点结构如图 5 所示。页面树形结构可表示为(html(head(meta(title，keywords，description)，style，script))，(body(table(tr(td(text))))，(div(ul(li(a))，(span(*text*)))，script…)。

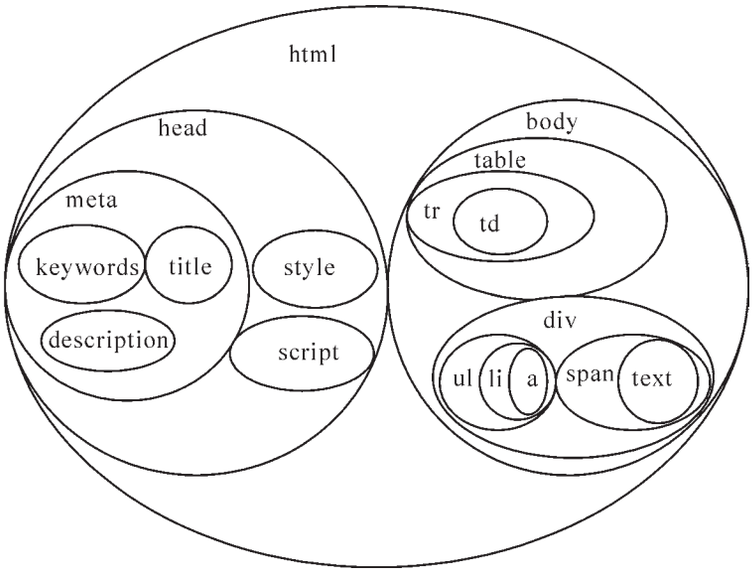

图 5　森林经营主题网页树形结构

Figure 5　Tree structure of forest management theme page

森林经营主题网页的视觉特征一般体现在字体、背景颜色、段落划分等方面；语义信息一般表现为页面内容的类型，如文本、多媒体或超链接等[11]。根据树形表示，能够发现森林经营主题网页的视觉特征节点和页面内容节点。head 部分的 title，keywords，description 是直接对页面或者森林经营知识的描述，与网页内容的符合度很高，可以用来辨别是否与森林经营主题相关。在森林经营主题网页中，style 和 script 部分体现的是视觉特征，对森林经营知识采集形成了干扰，全部当做噪声处理。森林经营知识一般隐藏在表格＜td＞＜/td＞，＜span＞＜/span＞，＜b＞＜/b＞，＜h1＞＜/h1＞等文本标记标签内，需要在匹配之后再进行提纯，获得知识。而＜ul＞＜/ul＞，＜li＞＜/li＞等标记可以用来定位地址并获得锚文本信息。锚文本信息计算辨别是否与森林经营知识主题相关后，可以判定地址是否加入爬行队列。除此之外的其他标记部分绝大多数都是对森林经营知识采集构成干扰的噪声。

3.2　经营知识智能匹配

森林经营知识采集需要进行知识抽取，从森林经营主题网页中包含的无结构或半结构信息中识别出与森林经营知识相关的数据，并转化为结构和语义更为清晰的格式[12]。在本研究中，采用基于 HTML 结构的方法实现森林经营知识抽取。这种抽取方法需要用正则表达式实现。通常，系统的正则表达式都是固定不变的，但是本研究中除了系统规则库中所包含的知识采集正则表达式外，还支持用户为特定页面指定 HTML 标签规则，并智能为其生成正则表达式。支持自定义规则的除了内容部分，还有页面属性(其中包括 title，keywords，description 等 meta 信息)，地址，文章标题，发布时间，信息来源等。自定义的规则使系统的匹配可以满足用户的多样性要求。爬虫需要对森林经营网页源码进行分析，获得页面包含的地址组，形成工作队列。假定抽取页面全部链接的正则表达式为 z_1，z_1 的表示如式(1)所示：

$$z_1=(?<=\text{href}*?=*?[\backslash'\backslash''])\text{.}*?(?=[\backslash'\backslash''])。\tag{1}$$

在根据用户输入的 URL 特征，进行特征 URL 抽取中，假设待生成的正则表达式为 z_2，z_2 表示如式(2)所示：

$$z_2=''\ (?<=''+s_1\text{.split}[i]+'')\text{.}*?(?=''+s_1\text{.split}[i+1]+'')''。\tag{2}$$

根据生成的正则表达式，系统会分析爬虫采集到的页面源码，然后抽取待采集的地址队列。正则表达式并非一成不变的，在定义的过程中考虑了多种情况进行处理。①假设森林经营知识的标题使用了＜h1＞＜/h1＞标记，但是页面中可能有多个相同标记，在此标记中，特征方法并不唯一，因此抽取全部的＜h1＞＜/h1＞标记内容，再进一步分析。在处理时，使用正则表达式的几种通配符，如(.+?)和(*)来进行通配处理。假设抽取标题内容的正则表达式为 z_3，z_3 表示如式(3)所示：

$$z_3=''(?<=''+s_2.\text{split}[i]+''+?>)*?(?=''+s_2.\text{split}[i+1]+'')''。 \quad (3)$$

在取得全部的标签队列之后，再分析队列各条记录的特征信息，如 id，class 等标记内容，根据不同网站的特征属性，提取森林经营知识信息。②相对于地址抽取和标题抽取来说，正文的抽取规则更复杂。正文抽取不但要求精准，还需要保留适当的换行格式等，方便直接在决策支持系统中应用。假设抽取正文的正则表达式为 z_4，z_4 表示如式(4)所示：

$$z_4=''(?<=''+s_3.\text{split}[i]+'')([^]*?)(?=''+s_3.\text{split}[i+1]+'')''。 \quad (4)$$

抽取获得正文后，再甄别内容，去掉与森林经营知识无关的视觉特征标签，保留部分换行标签，如＜*hr*/＞，＜*br*/＞等。在式(2)式(3)式(4)中，s_1，s_2，s_3 均表示用户输入的标签字符串，split[i]表示字符串被 split( )方法切分后的字符串数组在下标为 i 处的值。

3.3 森林经营知识去重

爬虫在采集之前，首先要查找地址库，查看是否已经采集了这条地址。如果没有重复，再进行页面采集。在经过主题过滤、抽取形成知识后，还要检测是否有重复的知识。知识重复检测的总体思想是为每个采集到的森林经营知识生成一个指纹，采用基于字符串比较的方法[13]计算 2 个指纹的相似性。若 2 个知识的指纹相似性大于某个阈值，则认为这 2 个知识重复。欧氏距离是空间中常用的计算 2 个 n 维向量距离的方法。欧式距离值越大，向量距离越远，文档相似程度越低；而欧式距离越小，向量距离越近，文档相似程度越高。将要对比的森林经营知识使用 TF-IDF 建立向量 V_q 和向量 V_d，再使用式(5)计算。

$$\text{sim}(q,\ d)=\text{sim}(V_q,\ V_d)=E_{Dqd}=\sqrt{\sum_{i=1}^{n}[(\text{t}q_i,\omega q_i)-(\text{t}d_i,\omega d_i)^2]}。 \quad (5)$$

求得的值越小，则知识相似度越高；值越大，则知识相似度越低。若值不在约定域内，将此页的地址和处理后的森林经营知识存入数据库。

4 实验分析

4.1 采集结果对比

本研究选择中国林业网的主站及其站群等权威网站作为主要测试网站，设定抓取时间为 1.0 h，共计采集链接样本数据 2.32 M，约有 2.5 万条链接数据。其中 2/3 的数据作为训练样本，1/3 的数据作为测试样本。爬虫将抓取的数据进行处理，形成结果(表 1)。实验结果表明：本系统采用的知识采集方式虽然在保存的总链接数目上少于普通爬虫方式，但采集到的符合森林经营知识主题的链接数比普通爬虫更多，符合主题的链接数所占百分比远远大于普通爬虫方式。

表 1 采集结果对比表

Table 1 Comparison table of acquisition results

方式	抓取链接数/个	保存链接数/个	符合主题数/个	符合主题数占抓取链接百分比/%
改进爬虫	12 785	6 377	6 377	49.87
普通爬虫	24 543	21 312	4 523	18.43

4.2 采集数据质量对比

本研究针对普通爬虫模拟工具抓取数据和森林经营知识采集系统所采集的知识数据质量进行了对比实验。2 种方式对抓取的链接地址进行处理后的结果分别如图 6 和图 7 所示(以“http://eucalypt.forestry.gov.cn/lzzmas/16043.jhtml”为例)。由结果可知：普通爬虫只对页面进行了粗略的去 HTML 标签处理，将页面所有文本保留，不进行精确的内容抽取和内容格式化处理，不能形成森林经营知识。

森林经营知识采集系统通过智能规则匹配对页面信息进行格式化处理和精确的内容处理，使得信息被分割成各个属性，包括知识标题、发布时间、知识来源、知识内容等方面。系统还对知识进行可信度评价，以直接存储到知识库，在森林经营决策支持系统中进行应用。

5 结论与讨论

知识丰富度决定了决策支持系统的问题处理能力。本研究研建的森林经营知识采集系统解决了在互

联网上获取森林经营知识的问题，提升了森林经营决策支持系统的知识丰富度。

本研究在分析森林经营知识采集问题的基础上，建立林业专有词库，改进网络爬虫规则，并利用森林经营主题爬虫算法、森林经营网页去噪、森林经营知识智能匹配、森林经营知识去重等技术，设计并实现了森林经营知识采集系统。本研究分析了森林经营主题网站的特点，建立了森林经营特征向量对采集内容进行过滤，使用欧氏距离进行森林经营知识指纹识别，获得了高相关度、高准确率、低重复度的森林经营知识。

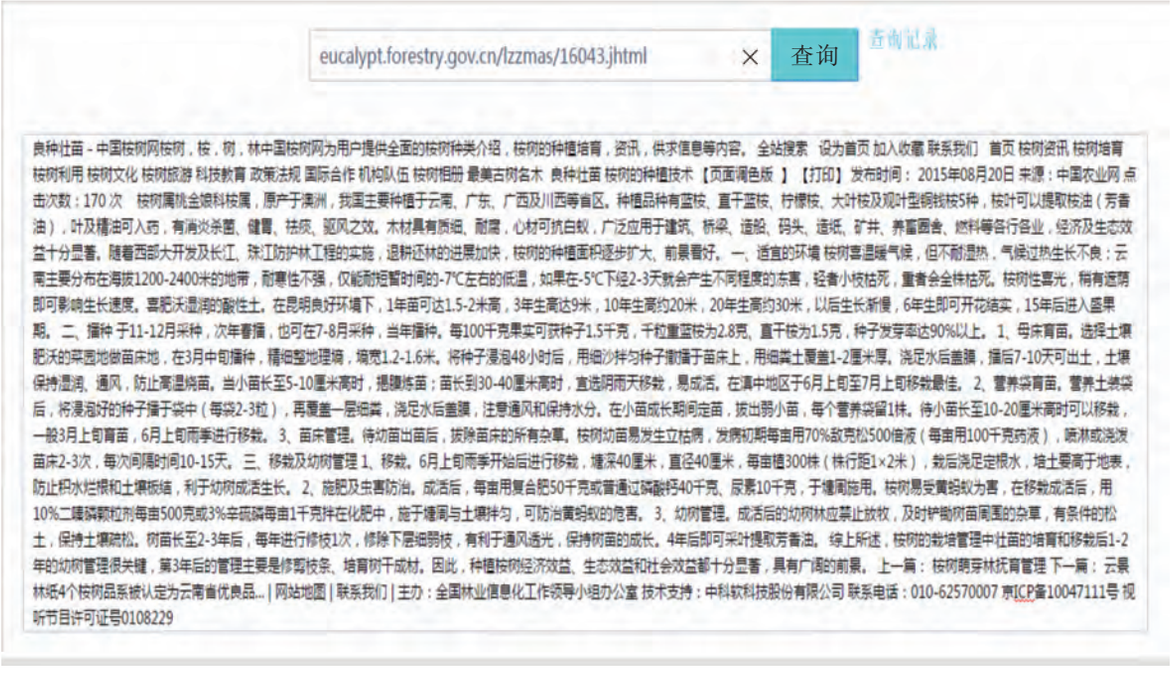

图 6　普通爬虫模拟工具抓取结果

Figure 6　Grab result of common crawler simulation tool

图 7　系统知识抽取结果

Figure 7　Knowledge extract result of the system

该系统已应用在国家高技术研究发展计划项目“数字化森林与牧场经营管理关键技术研究”中，长期为森林经营决策支持系统提供知识采集服务。

6 参考文献

[1] 吴保国，李成赞，马驰，等. 森林培育专家决策支持系统的研究[J]. 北京林业大学学报，2009，**31**(增刊 2)：1 – 8.
WU Baoguo, LI Chengzan, MA Chi, *et al*. An expert decision support system for silviculture [J]. *J Beijing For Univ*, 2009, **31**(supp 2): 1 – 8.

[2] 张戬慧. 专业智能搜索系统在动物医学领域中的应用[J]. 东北农业大学学报，2009，**40**(9)：141 – 144.
ZHANG Jianhui. Application of professional intelligent search system in veterinary medicine [J]. *J Northeast Agric Univ*, 2009, **40**(9): 141 – 144.

[3] 申晋. 基于 Lucene 和 Nutch 的林业垂直搜索引擎的研建[J]. 农业网络信息，2008(4)：16 – 18.
SHEN Jin. Study and implementation of forest vertical search engine based on Lucene and Nutch [J]. *Agric Network Inf*, 2008(4): 16 – 18.

[4] 袁津生，郭艳芬. 林业主题爬虫的算法研究与设计[J]. 计算机工程与设计，2011，**32**(6)：2003 – 2006.
YUAN Jinsheng, GUO Yanfen. Algorithm research and design of forestry focused web crawler [J]. *Comput Eng Des*, 2011, **32**(6): 2003 – 2006.

[5] 张丽莎，张贵，龙朝夕，等. 林业专题动态信息的搜索与集成[J]. 中南林业科技大学学报，2013，**33**(5)：47 – 51.
ZHANG Lisha, ZHANG Gui, LONG Chaoxi, *et al*. Search and integration of thematic dynamic information on forestry [J]. *J Cent South Univ For Technol*, 2013, **33**(5): 47 – 51.

[6] 李嘉，徐前，王梓，等. 基于语义的林产品贸易 Web 信息抽取算法[J]. 计算机工程与应用，2014，**50**(19)：199 – 204.
LI Jia, XU Qian, WANG Zi, *et al*. Forest products trading Web messages extraction algorithm based on semantic [J]. *Comput Eng Appl*, 2014, **50**(19): 199 – 204.

[7] 邓厚平，武刚. 基于爬虫和网站分类的主题信息源发现方法[J]. 计算机工程与应用，2016，**52**(3)：59 – 65.
DENG Houping, WU Gang. Discovery of topic-specific information source based on web crawler and website classification [J]. *Comput Eng Appl*, 2016, **52**(3): 59 – 65.

[8] 刘金红，陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究，2007，**24**(10)：26 – 29.
LIU Jinhong, LU Yuliang. Survey on topic-focused Web crawler [J]. *Appl Res Comput*, 2007, **24**(10): 26 – 29.

[9] 王娟，吴金鹏. 网络爬虫的设计与实现[J]. 软件导刊，2012，**11**(4)：136 – 137.
WANG Juan, WU Jinpeng. The design and implementation of Web crawler [J]. *Software Guide*, 2012, **11**(4): 136 – 137.

[10] 龚炳江，黄彦欣，贾海鑫. 矿山设备领域主题爬虫研究与设计[J]. 计算机应用与软件，2014，**31**(11)：122 – 124.
GONG Bingjiang, HUANG Yanxin, JIA Haixin. Studying and designing topic crawler for mining equipments field [J]. *Comput Appl Software*, 2014, **31**(11): 122 – 124.

[11] 丁宝琼，谢远平，吴琼. 基于改进 DOM 树的网页去噪声方法[J]. 计算机应用，2009，**29**(增刊 1)：175 – 177.
DING Baoqiong, XIE Yuanping, WU Qiong. Noise elimination method in Web page based on improved DOM tree [J]. *J Comput Appl*, 2009, **29**(supp 1): 175 – 177.

[12] 金岳富，范剑英，冯扬. 分布式 Web 信息采集系统的设计与实现[J]. 哈尔滨理工大学学报，2010，**15**(1)：116 – 119.
JIN Yuefu, FAN Jianying, FENG Yang. Design and realization of distributed Web crawler [J]. *J Harbin Univ Sci Technol*, 2010, **15**(1): 116 – 119.

[13] 秦杰，闫付亮，朱海丰，等. 基于链接信息的网页分类算法[J]. 微电子学与计算机，2012，**29**(6)：108 – 112.
QIN Jie, YAN Fuliang, ZHU Haifeng, *et al*. A webpage classification algorithm based on link information [J]. *Microelectron Comput*, 2012, **29**(6): 108 – 112.